The quantitative proteome of a human cell line

Martin Beck, Alexander Schmidt, Johan Malmstrom, Manfred Claassen, Alessandro Ori, Anna Szymborska, Franz Herzog, Oliver Rinner, Jan Ellenberg, Ruedi Aebersold

Corresponding author: Ruedi Aebersold, ETH Zurich

Review timeline:

Submission date:	15 August 2011
Editorial Decision:	16 September 2011
Revision received:	22 September 2011
Accepted:	29 September 2011

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision 16 September 2011

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from the two referees who accepted to evaluate the study. As you will see, the referees find the topic of your study of potential interest and are supportive. They raise however a series of concerns and make suggestions for modifications, which we would ask you to carefully address in a revision of the present work.

Please include a Materials and Methods section in the main manuscript. Since the current description of the methods is very detailed, which we greatly appreciate, it may not fit in the main paper. We would thus suggest to include a note stating that the full methods are described in Suppl information and include at least a "data availability section". Please deposit your raw data in one of the major public databases and include the respective hashcode/accession in this section.

*** PLEASE NOTE *** As part of the EMBO Publications transparent editorial process initiative (see our Editorial at http://www.nature.com/msb/journal/v6/n1/full/msb201072.html), Molecular Systems Biology will publish online a Review Process File to accompany accepted manuscripts. When preparing your letter of response, please be aware that in the event of acceptance, your cover letter/point-by-point document will be included as part of this File, which will be available to the scientific community. More information about this initiative is available in our Instructions to Authors.

Thank you for submitting this paper to Molecular Systems Biology.

Yours sincerely,

Editor
Molecular Systems Biology

Referee reports

Reviewer #1 (Remarks to the Author):

Review for Beck/Schmidt et al. - The quantitative proteome of a human cell line.

This is one of two manuscripts submitted back-to-back to Molecular Systems Biology using an established human cancer cell line model to obtain an in-depth proteome description.

Both manuscripts are of high quality and impressively demonstrate the capabilities of modern shotgun proteomics to basically identify, and to a certain degree quantify, the proteome of a mammalian cell. Both manuscripts arrive at similar results and conclusion, despite using slightly different mining strategies. Although, not geared towards specific biology, both manuscripts provide the first true high-level cell/systems level insight into the nearly complete proteome of a simple model system. These data will provide the cornerstone for similar proteome projects in the future.

The data quality of both papers is very high and the high-level analysis of this large amount of data is useful and reasonable. Asides from "real" biology, which is clearly not the focus of these current manuscripts there are only a few minor comments that should be addressed editorially.

The quantification provided by Beck/Schmidt is likely more accurate, compared to the second paper, since isotope spike-in experiments were performed. Also the comparison to fluorescence microscopy of the nuclear pore complex was quite convincing.

1) Raw data should be deposited to Tranche. This will enable further mining of these data by others.
2) Page 16: ...naÔve protein FDR estimate. What is this? A simple target-decoy search?
3) The colors in Figure 3 are very difficult to distinguish. Somehow this figure needs to be rearranged.

In summary, both papers are highly suitable for publication in MSB and should be of high interest to the systems biology readership.

Reviewer #2 (Remarks to the Author):

The impressive work described by the manuscript from Beck and colleagues presents an important (technical) advancement in the study of human proteomes and systems biology. As such, it should (eventually) be published in MSB despite its very descriptive character and the lack of any follow-up experiments. Before the manuscript is accepted for publication, I strongly suggest some additional comments and discussion about several technical aspects which are very short or omitted in the current version. I don't mean to be nitpicking, but I think it adds to the paper to discuss also some of the limitations of the chosen strategies as this will allow the community to built on this resource, and revise and improve it.

The full proteome was assembled from MS measurements of different OGE fractions. The variations in the composition of these fractions, and specifically the variable ionizability and amounts of the contained peptides (based on isoelectric point for example) can affect the extracted ion currents in batch-specific ways, with some fractions containing peptides that are overall more ionizable and produce higher XICs than those in other fractions. How did the analysis adjust for these batch effects? The methods describe a "union [that] took into account the extracted ion currents observed for each peptide species at all possible charge states in all OGE fractions," but the manuscript does not describe what fraction of peptides/proteins were normalized across different fractions in this way. Correction or explanation of possible batch effects would strengthen the confidence of the reported quantifications (see e.g. Biostatistics 8:118, 2007).

The extent of the measured proteome is impressive, but is of course not exhaustive (despite the authors' repeated claim). What gene products that have been measured to be expressed in previous

studies of U2OS cells were not detected by the reported approach? Gene expression in U2OS cells has been extensively studied using oligonucleotide microarrays, as available from either the NCBI Gene Expression Omnibus or the Cancer Cell Line Encyclopedia http://www.broadinstitute.org/ccle/home. Specifically, what fraction of genes that were found to be highly expressed at the mRNA level in previously published studies were not detected using rolling inclusion lists? This information will be useful in understanding the limitations of the used approach and potential ways to improve it.

What is the origin of U2OS cells used in this study? Since cancer cell lines can vary significantly from clone to clone, particularly for chromosomally unstable cell lines such as U2OS, such information will be crucial for future experiments by others should they wish to carry out comparisons with the results presented in the submitted manuscript. At the very least, a revised manuscript should describe the detailed source of the actual cell line used for the study (so that it may be obtained by others). If possible, the authors may choose to provide genotyping information, as is becoming more commonplace: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-017.html.

Along similar lines: it is well known for cancer cell lines that in order to proliferate so rapidly (hence cancerous) they generate a lot of ATP using the glycolysis pathway. In that sense, it is not surprising that the proteome of this cell line is enriched for protein complexes representing glycolysis. However, the authors have compared their findings to a mouse cell line, NIH3T3, which is a fibroblast cell line. I find it surprising that the results are very similar. Could the authors comment on that i.e. if there could be a comparison of the cancer-related genes between these cell lines and whether some of the proteins in the glycolysis pathway are over-represented in the osteosarcoma cell line?

Please provide a list of the 144 peptides and 79 proteins covered by their proteotypic test peptides. Are the localization and functional classes of these proteins similar to the proteins identified? How many membrane/cytosolic proteins were covered by this set of 79 proteins? How well did the intensities of these 144 peptide correlate with their nominal concentration? Given the fact that the authors claim to have "successfully determined the copy numbers per cell for 73% of U2OS proteins that are detectable with the MS method used, with an estimated mean error of about 2-fold" it would be useful to see the correlation function of the 144 model peptides and whether they all lie on the same line with a less than 2-fold mean error.

It would be helpful if the authors were to discuss the quantification approach a little more critically. At this moment, their proteotypic peptides cover less than three orders of magnitude ranging from 2,500,000 to 4,500 copy numbers. Nonchalantly, they extended their linear range by one order of magnitude in each direction to 20,000,000 and 500. This might be a perfectly justifiable approach given the extent of their work, but it might be worthwhile to clearly list the assumptions made for their absolute quantification.
Secondly, it should be highlighted that the authors at this moment confirmed a very narrow range of 0.2 log units compared to the 3 orders covered by their calibration curve. Again, given the size of this study this might be justifiable, but should be listed as one of the assumptions.
Thirdly, on page 5 the authors state "that we successfully determined the copy numbers per cell for 73% of U2OS proteins that are detectable with the MS method used, with an estimated mean error of about 2-fold." One could argue that even the box/whisker plot in Figure 2b supports this statement. However, based on this box/whisker plot, one could argue that this statement only holds true for 50% of the proteins (the boxes), whereas the other 50 % (whiskers and beyond) cover several orders of magnitude, i.e. corresponding to a much larger error.

On page 7 the authors state "However, oxidative phosphorylation and electron transport are for the most part catalyzed by proteins of moderate abundance,... ." This might be an artifact of underrepresented membrane proteins, an issue the authors have not really addressed, as both classes of proteins are primarily found in the mitochondrial membranes. One could argue that this point is corroborated by the subsequent statement "Secondly, we found that cellular processes associated with protein synthesis and turn-over, namely translation, protein folding, splicing and degradation as well as RNA processing, are mostly conducted by high abundant proteins." All the proteins listed are well detectable cytosolic proteins.

On page 11, the authors state "In particular, intracellular proteins were overrepresented on the cost of membrane proteins on the proteome level (Table S4). Such an effect has been observed before (Schrimpf et al, 2009) and is likely a result of the reduced accessibility of membrane proteins for MS analysis, although we had used an MS compatible detergent during sample preparation." It is not clear to me how the authors addressed (and tested) this issue in their study. How did they ensure (and proved their point) that they have an equal representation of membrane and cytosolic proteins?

Minor comments:

Some of the figures use font sizes that will be too small to be legible once reduced to print size (for example, Fig. 1b, 2c).

Though previous proteomics studies of U2OS cells used an inferior approach (Cancer Genomics Proteomics 5:63, 2008), this published work should be cited in the manuscript.

Figure 2d: It would be helpful if x-axis label was revised to clarify that the authors talk about frequency of occurrence.

On page 2, there is a comment ("should we name the %") which seems to be for internal use only. Please delete and ensure that there are no more instances. And to answer the question: yes, it would be helpful if the names of the two bacteria types were added.

At the top of page 5, the authors write "We used the same data to calculate three different protein abundance scores as previously described and validated their precision by statistical analysis (Figure S1)." Please provide details and/or references for the relevant previously described protein abundance scores.

1st Revision - authors' response 22 September 2011

We thank both reviewers for their constructive criticism that largely contributed to improving our manuscript during the revisions. Our detailed point by point reply is listed below:

Reviewer #1
"Review for Beck/Schmidt et al. - The quantitative proteome of a human cell line.

This is one of two manuscripts submitted back-to-back to Molecular Systems Biology using an established human cancer cell line model to obtain an in-depth proteome description.

Both manuscripts are of high quality and impressively demonstrate the capabilities of modern shotgun proteomics to basically identify, and to a certain degree quantify, the proteome of a mammalian cell. Both manuscripts arrive at similar results and conclusion, despite using slightly different mining strategies. Although, not geared towards specific biology, both manuscripts provide the first true high-level cell/systems level insight into the nearly complete proteome of a simple model system. These data will provide the cornerstone for similar proteome projects in the future.

The data quality of both papers is very high and the high-level analysis of this large amount of data is useful and reasonable. Asides from "real" biology, which is clearly not the focus of these current manuscripts there are only a few minor comments that should be addressed editorially.

The quantification provided by Beck/Schmidt is likely more accurate, compared to the second paper, since isotope spike-in experiments were performed. Also the comparison to fluorescence microscopy of the nuclear pore complex was quite convincing.

1) Raw data should be deposited to Tranche. This will enable further mining of these data by others."

As requested we have deposited the data. The following three hash codes corresponding to the proteome mapping and both quantification experiments are listed in the manuscript:

3Wj0424JA2DCVkBnfqm45v+UfMZOHgf3p2PTwUe83RwjqQvtr4mQnloYvUSrMHCYBz+krDIXmz50s
pF2TNYGw3/8jZIAAAAAAAAB9w==

x8hmYUs40bOspaY+EMuyUtDkyiw+xgyjSynVK/ggQXhl+bbDV5QbiAMakzsSKonz/XszxEEUThtmn
6cIS/STS1Y0n2QAAAAAAAAB3g==

Gm5TsXK3crQV70MqiIIH+/uaKyioNCFWi+Ri7fpLq+W1ga5OQA0dTe2u0LMvN+ty7uuRsA1o3WTWb
79Bc/XqYK7v9D0AAAAAAAACBA==

"2) Page 16: ...naive protein FDR estimate. What is this? A simple target-decoy search?"

The naive estimate computes protein FDR directly from the number of decoy peptide identifications, thereby neglecting the situation of true protein identifications being both supported by true and false peptide spectrum matches. The consequence of this is that for large datasets like the ones generated in this study the number of false positive proteins explodes if this problem is not controlled. The Mayu software tool corrects for this

phenomenon by appropriately modeling its statistical implications on the FDR estimate. We have revised the respective figure legend for more clarity.

"3) The colors in Figure 3 are very difficult to distinguish. Somehow this figure needs to be rearranged."

We discovered that the color coding in this figure was represented incorrectly in the legend, which has most probably caused the confusion. We have corrected for this error in the revised version of the manuscript and inserted separators between the general categories to improve the visualization.

"In summary, both papers are highly suitable for publication in MSB and should be of high interest to the systems biology readership."

Reviewer #2

"The impressive work described by the manuscript from Beck and colleagues presents an important (technical) advancement in the study of human proteomes and systems biology. As such, it should (eventually) be published in MSB despite its very descriptive character and the lack of any follow-up experiments. Before the manuscript is accepted for publication, I strongly suggest some additional comments and discussion about several technical aspects which are very short or omitted in the current version. I don't mean to be nitpicking, but I think it adds to the paper to discuss also some of the limitations of the chosen strategies as this will allow the community to built on this resource, and revise and improve it.

The full proteome was assembled from MS measurements of different OGE fractions. The variations in the composition of these fractions, and specifically the variable ionizability and amounts of the contained peptides (based on isoelectric point for example) can affect the extracted ion currents in batch-specific ways, with some fractions containing peptides that are overall more ionizable and produce higher XICs than those in other fractions. How did the analysis adjust for these batch effects? The methods describe a "union [that] took into account the extracted ion currents observed for each peptide species at all possible charge states in all OGE fractions," but the manuscript does not describe what fraction of peptides/proteins were normalized across different fractions in this way. Correction or explanation of possible batch effects would strengthen the confidence of the reported quantifications (see e.g. Biostatistics 8:118, 2007)."

We first want to point out that this effect, correctly described by the reviewer, is not affecting our proteome mapping experiment and, more importantly, is not expected to have a significant impact onto the absolute quantification estimates presented here. This is because spectral counts were used instead off XICs. Nevertheless, we agree with the reviewer that generating extracted XICs from fractionated samples might be affected by various local sample complexities and therefore vary across fractions, although we believe that this effect is rather small: In experiments that compare peptide XICs of non-fractionated and fractionated samples, we do see a very good correlation of intensities, indicating that peptide ionization is affected only for a minority of features (see e.g. Schmidt et al, MSB 2011). Therefore, the sum of peptide intensities detected across all fractions represents a good intensity measure for most identified peptides. The method

this reviewer refers to (Biostatistics 8:118, 2007) was established for microarrays. As far as the flyability of peptides is concerned, the prediction of ionizibilty in different (peptide compositional) backgrounds is not straight forward, since no mathematical models for this phenomenon exist. However, several classification tools have been developed, including *PeptideSieve* from our group, that address the issue using empirical data. To our knowledge a dependency on the isoelectric point was not reported so far. It is thus not clear to us on which bases one should carry out the normalization without introducing quantification artifacts by 'raising' or 'lowering' individual fractions. Since we pooled low complexity fractions, the sample complexity was very similar across fractions. We had thus decided not to normalize for such effects since it is very difficult. Furthermore, the reference peptides were spiked into the samples before fractionation and should thus be unaffected in terms of quantification.

"The extent of the measured proteome is impressive, but is of course not exhaustive (despite the authors' repeated claim). What gene products that have been measured to be expressed in previous studies of U2OS cells were not detected by the reported approach? Gene expression in U2OS cells has been extensively studied using oligonucleotide microarrays, as available from either the NCBI Gene Expression Omnibus or the Cancer Cell Line Encyclopedia http://www.broadinstitute.org/ccle/home. Specifically, what fraction of genes that were found to be highly expressed at the mRNA level in previously published studies were not detected using rolling inclusion lists? This information will be useful in understanding the limitations of the used approach and potential ways to improve it."

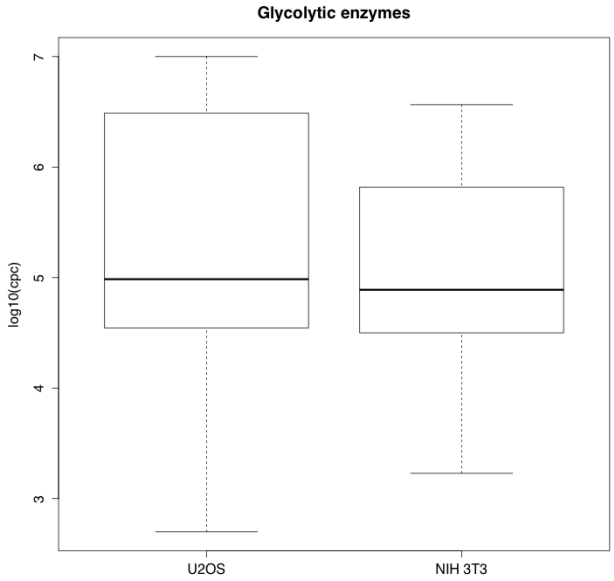
This was a very good suggestion! We have done this analysis and it further underlines our finding that membrane proteins are, at least to some extent, less accessible for the proteomic approach used in this study and for that matter, any other proteomic method (please see also response to other points below). We added the following text to the section of the results part dealing with proteome completeness: 'To assess the comprehensiveness of our approach we asked whether mRNAs that are highly expressed in u2os cells remain undetected on the protein level. Based on RKPM values (reads per kilobase of exon model per million mapped reads) provided by the aforementioned study of u2os cells (Lundberg et al, 2010), we detected proteins corresponding to ~84% of the most abundant quartile of mRNAs. Although it remains unknown whether all mRNAs are translated into proteins, gene ontology analysis revealed that ~ 33% of the mRNAs unidentified on the protein level encode transmembrane proteins, suggesting that these proteins are less accessible for our proteomic approach (discussed below).', and have adjusted the discussion part accordingly.

"What is the origin of U2OS cells used in this study? Since cancer cell lines can vary significantly from clone to clone, particularly for chromosomally unstable cell lines such as U2OS, such information will be crucial for future experiments by others should they wish to carry out comparisons with the results presented in the submitted manuscript. At the very least, a revised manuscript should describe the detailed source of the actual cell line used for the study (so that it may be obtained by others). If possible, the authors may choose to provide genotyping information, as is becoming more commonplace: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-017.html."

The cells were originally obtained from the American Tissue Type Culture collection. We included to ordering information into the supplement of the revised manuscript. At this point genotyping information is not available to us.

"Along similar lines: it is well known for cancer cell lines that in order to proliferate so rapidly (hence cancerous) they generate a lot of ATP using the glycolysis pathway. In that sense, it is not surprising that the proteome of this cell line is enriched for protein complexes representing glycolysis. However, the authors have compared their findings to a mouse cell line, NIH3T3, which is a fibroblast cell line. I find it surprising that the results are very similar. Could the authors comment on that i.e. if there could be a comparison of the cancer-related genes between these cell lines and whether some of the proteins in the glycolysis pathway are over-represented in the osteosarcoma cell line?"

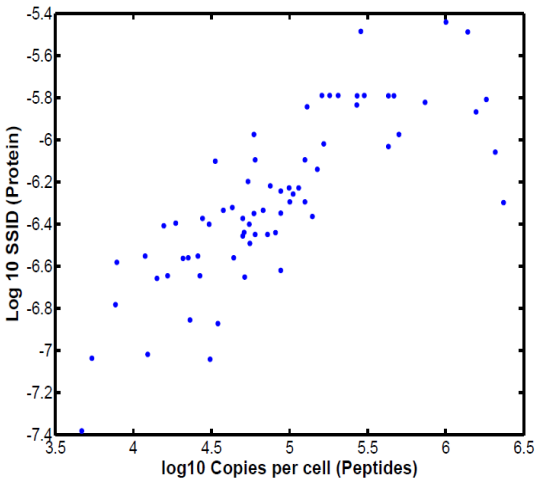
We indeed find glycolytic enzymes to be expressed at high number in u2os (median copy number of 1e5) and, therefore, contributing to a significant portion of the proteome of these cells. Regarding the comparison with NIH-3T3 fibroblast, a closer investigation in the core glycolytic enzymes (see figure below) confirms a similar expression profile of these proteins in the two cell lines. Moreover, despite its cancerous nature, U2OS displays a doubling time that is significantly longer than NIH-3T3: ~30 h for U2OS (Russo, Magro, *et al.* Cancer Research (2006) against ~20 h for NIH-3T3 (DSMZ – the German Resource Centre for Biological Material, www.dsmz.de). We would therefore expect a higher metabolic rate in this fibroblast cell line than in U2OS. Finally, we are not aware whether the solely protein expression level of the glycolytic enzymes can be a sufficient predictor of the metabolic state of a cell.

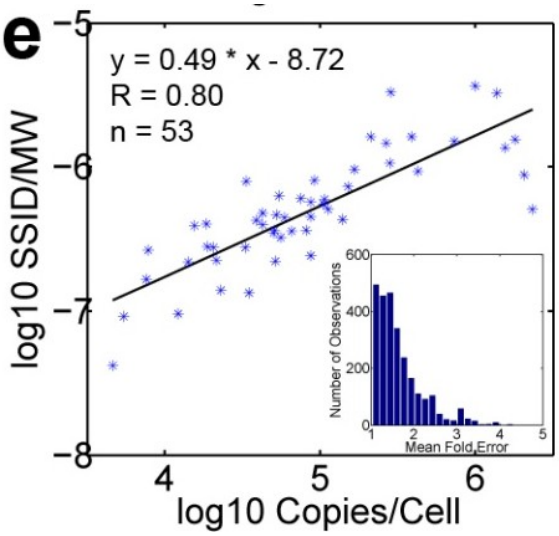


"Please provide a list of the 144 peptides and 79 proteins covered by their proteotypic test peptides. Are the localization and functional classes of these proteins similar to the proteins identified? How many membrane/cytosolic proteins were covered by this set of 79 proteins? How well

did the intensities of these 144 peptide correlate with their nominal concentration? Given the fact that the authors claim to have "successfully determined the copy numbers per cell for 73% of U2OS proteins that are detectable with the MS method used, with an estimated mean error of about 2-fold" it would be useful to see the correlation function of the 144 model peptides and whether they all lie on the same line with a less than 2-fold mean error."

We have included a list of the peptides as requested (Table S1). We realized that the set of peptides spiked in to samples accounted for 84 instead of 79 proteins. We apologize for this error and have corrected for this in the revised version of the manuscript. This has no effect on the number of detected peptides (70 peptides corresponding to 53 proteins). In contrast the proteome map proteins of which we selected the peptides for quantification do not seem to show a significant over or under representation of any functional class. Out of the 84 proteins, 54 are annotated as cytoplasmic and 7 as integral to membrane. Such annotations are of course neither unique nor comprehensive. The correlation on peptide level is very similar to the protein level, as one can see in the plots inserted below (left: peptide level correlation, y = 0.5011x - 8.7418, R = 0.79; right: snapshot of Figure S1e - correlation on protein level). Since the global protein quantification is calculated on the protein and not peptide level, we found this information to be less relevant then the data displayed in Figure S1.





"It would be helpful if the authors were to discuss the quantification approach a little more critically. At this moment, their proteotypic peptides cover less than three orders of magnitude ranging from 2,500,000 to 4,500 copy numbers. Nonchalantly, they extended their linear range by one order of magnitude in each direction to 20,000,000 and 500. This might be a perfectly justifiable approach given the extent of their work, but it might be worthwhile to clearly list the assumptions made for their absolute quantification."

We appreciate this comment. In response we like to point out that the issues are clearly addressed in the main text within the same paragraph and context, namely the fact that we validated within 'a concentration range from 4.5exp3 to 2.5exp6 copy numbers per cell' and as well that 'Since we cannot assess

the precision of the quantitative estimates outside of the dynamic range covered by heavy labeled reference peptides, we masked proteins below 500 copies per cell and above 20,000,000 copies per cell, respectively'. We also believe that we went much further with validation experiments than anyone before. This was explicitly appreciated by reviewer one (please see above). In order to 'list the assumptions made for their absolute quantification' more clearly, we revised the respective sentence to: 'Since we cannot assess the precision of the quantitative estimates outside of the dynamic range covered by heavy labeled reference peptides, we masked proteins below 500 copies per cell and above 20,000,000 copies per cell, respectively, assuming a correlation similar to the validated concentration range within the respective range of protein copies. '.

"Secondly, it should be highlighted that the authors at this moment confirmed a very narrow range of 0.2 log units compared to the 3 orders covered by their calibration curve. Again, given the size of this study this might be justifiable, but should be listed as one of the assumptions."

We assume that the reviewer refers to the measurement of the number of NPCs per cell by light microscopy. This analysis provides a single number, namely the average number of NPCs per cell together with its standard deviation accounting for the variation of this number across single cells. MS-derived copy numbers account for the average copy number measured in the bulk of lysate. Thereby copy numbers measured for individual nucleoporins are integrated into a single number based on known copy numbers per NPC. The precision indicated in the plot refers to the precision of the method as determined by bootstrap analysis. This is clearly stated in the figure legend. The comparison of both numbers serves for precisely one purpose namely to ensure 'that we neither systematically over- nor under-estimate protein abundance in the MS derived quantitative scale', as stated in the manuscript. We have modified the respective sentence in the main text to make this more transparent: 'Although this validation method relies only onto a single measurable value, namely NPC copies per cell, it demonstrates that we neither systematically over- nor under-estimate protein abundance in the MS derived quantitative scale.'.

"Thirdly, on page 5 the authors state "that we successfully determined the copy numbers per cell for 73% of U2OS proteins that are detectable with the MS method used, with an estimated mean error of about 2-fold." One could argue that even the box/whisker plot in Figure 2b supports this statement. However, based on this box/whisker plot, one could argue that this statement only holds true for 50% of the proteins (the boxes), whereas the other 50 % (whiskers and beyond) cover several orders of magnitude, i.e. corresponding to a much larger error."

We would like to point out that the estimated mean error for protein quantification and the boxes/whiskers in Figure 2b cannot be directly related to each other since they refer to different classes: single proteins in the first case and protein groups in the second. In particular, the box plots depict the distributions of protein copy number for different GO categories. These distributions are therefore representative of their protein group, they provide a global view of the expression range for a given category, and their broadness mostly reflect the heterogeneity of the grouped proteins. For instance, not all the kinases are expressed at the same level (for example MAPK-1 is expressed at high copy number 3.2e5, while upstream activators in

the same pathway are expressed at significant lower copy number, e.g. MAP4K-5 expressed at 1.7e3 copies and MAP4K-2 present at less than 1,000 copies). We maintain that the accuracy of the protein measurements is as stated.

"On page 7 the authors state "However, oxidative phosphorylation and electron transport are for the most part catalyzed by proteins of moderate abundance,... ." This might be an artifact of underrepresented membrane proteins, an issue the authors have not really addressed, as both classes of proteins are primarily found in the mitochondrial membranes. One could argue that this point is corroborated by the subsequent statement "Secondly, we found that cellular processes associated with protein synthesis and turn-over, namely translation, protein folding, splicing and degradation as well as RNA processing, are mostly conducted by high abundant proteins." All the proteins listed are well detectable cytosolic proteins."

That is a fair criticism! Our comment on the abundance of proteins involved in 'oxidative phosphorylation and electron transport' is not valid because these processes are mostly conducted by membrane proteins and this observation IS likely biased by the fact that membrane proteins were under-represented in general. We therefore have removed the corresponding sentence from the main text and adjusted the discussion correspondingly.

"On page 11, the authors state "In particular, intracellular proteins were overrepresented on the cost of membrane proteins on the proteome level (Table S4). Such an effect has been observed before (Schrimpf et al, 2009) and is likely a result of the reduced accessibility of membrane proteins for MS analysis, although we had used an MS compatible detergent during sample preparation." It is not clear to me how the authors addressed (and tested) this issue in their study. How did they ensure (and proved their point) that they have an equal representation of membrane and cytosolic proteins?"

We agree with the reviewer that our sentence was not clear enough. We therefore revised the manuscript as follows: "In particular, GO analysis reveals an under-representation of transmembrane proteins in the identified proteome (Table S4)." Table S4 reports the output of the GO analysis that we performed by comparing the identified proteome against the reference database (as stated in Supplementary Methods). The aim of this analysis was to point out eventual biases in the identification of certain protein categories. Transmembrane proteins are the major category that is under-represented in our work, as discussed in the manuscript.

Minor comments:

'Some of the figures use font sizes that will be too small to be legible once reduced to print size (for example, Fig. 1b, 2c).'

We corrected for this.

'Though previous proteomics studies of U2OS cells used an inferior approach (Cancer Genomics Proteomics 5:63, 2008), this published work should be cited in the manuscript.'

To our knowledge the most comprehensive study of u2os cells so far, is Lundberg et al., MSB 2010, which discovered 5399 proteins and was referred to in the main text, although not specifically mentioning u2os but mammalian cells. The study this reviewer mentions has discovered 237 proteins after 2DE and MS identification. This is only a fraction of the proteins discovered by Lundberg and coauthors. In this context we realized that Lundberg et al. and Niforou et al. had used different databases for searching their MS data then we have in our proteome mapping experiment. Since the different databases contain a different number of entries, these studies are not comparable to our work in terms of percentage of proteome coverage. To make the data obtained in the different studies better comparable to each other we decided to refer to absolute numbers instead of percentages in our revised manuscript and changed the respective passage as follows: 'From the identified peptides we inferred 10,006 proteins (Table S1, raw data available at https://proteomecommons.org), which is to our knowledge the by far the most comprehensive proteome map of a mammalian cell line, with earlier studies reaching e.g. 5399 proteins in u2os (Lundberg et al, 2010) and 2859 proteins in HeLa cells (Wisniewski et al, 2009). Earlier studies of u2os cells that used other proteomic approaches discovered even fewer proteins (n=237) (Niforou et al, 2008).' To make this more accessible for the reader we also clearly refer to the databases in the supplement, including version numbers.

'Figure 2d: It would be helpful if x-axis label was revised to clarify that the authors talk about frequency of occurrence.'

This issue has been addressed in the revised manuscript.

'On page 2, there is a comment ("should we name the %") which seems to be for internal use only. Please delete and ensure that there are no more instances. And to answer the question: yes, it would be helpful if the names of the two bacteria types were added.'

This issue has been addressed in the revised manuscript.

'At the top of page 5, the authors write "We used the same data to calculate three different protein abundance scores as previously described and validated their precision by statistical analysis (Figure S1)." Please provide details and/or references for the relevant previously described protein abundance scores.'

In the interest of readability we had removed all the technical details about protein abundance scores in an earlier version of the manuscript. The supplement, however, provides a detailed account on this issue. We agree with the reviewer that at least clear references to original work and to the supplement should be made in the main text and have revised it accordingly.